Question 1
Consider the total purchase cost of each product category and the statistical description of the dataset above for your sample customers. What kind of establishment (customer) could each of the three samples you've chosen represent?

Answer:
For Sample 0, the value for “Fresh” is far above the mean for the dataset. The remaining categories are significantly below the mean values. This could represent a grocery store or other establishment that sells primarily fresh produce.

Samples 1 and 2 have a “Fresh” value far below the mean, but the “Grocery” values are above the mean. Sample 2 has “Milk” and “Frozen” values below the mean, and Sample 1 has a “Detergents_Paper” value that is significantly above the mean. All three sample points have “Delicatessen” values below the mean.

These establishments could be restaurants that use large amounts of packaged food and use more paper products and detergent items to keep their establishments clean. The differences in the values between Samples 1 and 2 could indicate different types of restaurants or similar establishments.

Question 2
Which feature did you attempt to predict? What was the reported prediction score? Is this feature is necessary for identifying customers' spending habits?

Hint: The coefficient of determination, R^2, is scored between 0 and 1, with 1 being a perfect fit. A negative R^2 implies the model fails to fit the data.

Answer:
I attempted to predict Grocery, which has an R^2 of about .60, and Detergents_Paper, which has an R^2 of about .73. I found that these two features could be predicted with a greater degree of accuracy than the rest of the features in the dataset. This indicates that they might not be as necessary in identifying customers because their value can be derived from the other features in the dataset. They could probably safely be dropped from the dataset.

Question 3

Are there any pairs of features which exhibit some degree of correlation? Does this confirm or deny your suspicions about the relevance of the feature you attempted to predict? How is the data for those features distributed?
Hint: Is the data normally distributed? Where do most of the data points lie?

Answer:
Grocery and Detergents_Paper show some correlation, where most of the other pairs of features do not appear to. Milk also shows some correlation with Detergents_Paper and Grocery. This suggests that these features might be more important than the others in identifying customer groups. For Grocery and Detergents_Paper, the points appear to extend upward into a rough linear pattern starting at the origin, with most of the points concentrated nearer the origin.

The data is not normally distributed but heavily skewed to the right. Most of the data points lie on the left side of the distribution as indicated by the curves displayed on the scatter matrix.

Question 4
Are there any data points considered outliers for more than one feature? Should these data points be removed from the dataset? If any data points were added to the outliers list to be removed, explain why.

Answer:
I found 6 data points that were considered outliers for more than one feature. Because I felt that these data points might be less representative of a specific cluster or might belong to a different underlying category, I removed them from the dataset.

Question 5
How much variance in the data is explained **in total** by the first and second principal component? What about the first four principal components? Using the visualization provided above, discuss what the first four dimensions best represent in terms of customer spending.

Hint: A positive increase in a specific dimension corresponds with an increase of the positive-weighted features and a decrease of the negative-weighted features. The rate of increase or decrease is based on the individual feature weights.

Answer:
72% of the data is explained by the first and second principal components. The first four components together explain about 93% of the data. This suggests that there are between two and four dimensions to the data that we should be visualizing.

An increase in the first PC dimension corresponds to a decrease in “Fresh” and “Frozen,” and an increase in “Milk,” “Grocery,” and “Detergents_Paper.” This dimension suggests customers exist along a spectrum of purchasing produce to sell directly versus purchasing and selling other prepared items. An increase in the second dimension corresponds to an increase in all features, which is largest in “Fresh,” “Frozen,” and “Delicatessen.” This suggests that it makes sense to divide customers into those who purchase large amounts of fresh and frozen produce and those who do not.

Question 6
What are the advantages to using a K-Means clustering algorithm? What are the advantages to using a Gaussian Mixture Model clustering algorithm? Given your observations about the wholesale customer data so far, which of the two algorithms will you use and why?

Answer:
K-Means is fast and efficient and works best when you have an idea of how many clusters you will have in your problem. Gaussian Mixture Models are a kind of soft clustering, meaning they can assign a data point to more than one potential cluster. They are useful when, for example, you may have clusters of different sizes, which K-Means is not well-suited to handling.

Because there appear to be two kinds of customer in this problem from the PCA results, it would make sense to use two clusters, so I will be using K-Means since my number of clusters will likely be known in advance.

Question 7
Report the silhouette score for several cluster numbers you tried. Of these, which number of clusters has the best silhouette score?

Answer:

2 clusters: 0.4263

3 clusters: 0.3940

4 clusters: 0.3325

5 clusters: 0.3534

Of these, choosing 2 clusters gives the best silhouette score.

Question 8
Consider the total purchase cost of each product category for the representative data points above, and reference the statistical description of the dataset at the beginning of this project. What set of establishments could each of the customer segments represent?
Hint: A customer who is assigned to 'Cluster X' should best identify with the establishments represented by the feature set of 'Segment X'.

Answer:
For Segment 0, all of the product categories have purchase costs below the mean. For Segment 1, the “Fresh,” “Frozen,” and “Delicatessen” values are below the mean, while the other categories are above the mean by similar amounts.

Segment 1 seems to represent customers who purchase large amounts of fresh and frozen produce and could most likely be grocery stores or supermarkets that would have large sections of their store dedicated to produce. Having the true centers be below the mean suggests that there could be outliers skewing the distribution of purchase amounts.

Question 9
For each sample point, which customer segment from **Question 8** best represents it? Are the predictions for each sample point consistent with this?

Answer:

For Sample 0, the value for “Fresh” is much higher than both the mean for the dataset and the centers of both clusters. “Milk,” “Grocery,” and “Delicatessen” are close to the Segment 0 values. “Frozen” and “Detergents_Paper” are also closer to Segment 0 than Segment 1 values.

For Sample 1, the values all match Segment 1 closely except for “Detergents_Paper,” which is still closer to Segment 1 than to Segment 0. For Sample 3, the value for the “Fresh” category is very low and does not match either segment, but the remaining values are close to Segment 1 values.

The first point was predicted to be in the “grocery store” type category and its prediction is Segment 1, which agrees with the original prediction. The other two points were predicted to be in the “restaurant” category, and they were both predicted to be in Segment 0, which also agrees with the original prediction.

Question 10
Companies often run A/B tests when making small changes to their products or services. If the wholesale distributor wanted to change its delivery service from 5 days a week to 3 days a week, how would you use the structure of the data to help them decide on a group of customers to test?
Hint: Would such a change in the delivery service affect all customers equally? How could the distributor identify who it affects the most?

Answer:
If a company were going to change its delivery service and use A/B testing to evaluate the results, it would need to choose similar groups of customers as both the control and testing groups in order to get meaningful results, which could be identified through this cluster analysis of underlying categories in the customer base.

When the distributor first tested out the new delivery system, the likely problem was that the customers being tested all fell into a category that was not negatively affected by the evening deliveries, and customers who would be negatively affected were not being tested because they fell into the other category.

To resolve this, it should do one test of Retailer customers and one test of HoReCa customers to see how the treatment and control groups in each category responded to changes in delivery schedules. The treatment and control groups in each category should be as similar as possible to each other in terms of purchases.

Question 11

Assume the wholesale distributor wanted to predict some other feature for each customer based on the purchasing information available. How could the wholesale distributor use the structure of the data to assist a supervised learning analysis?

Answer:
Besides the six product features, a new feature that labels each customer as a Retailer or HoReCa customer could be added to the dataset. This could be used to predict new features such as whether the customer would accept evening delivery times, the actual results being determined after A/B testing and provided to train a supervised learning model.

Question 12 (Final Question)
How well does the clustering algorithm and number of clusters you've chosen compare to this underlying distribution of Hotel/Restaurant/Cafe customers to Retailer customers? Are there customer segments that would be classified as purely 'Retailers' or 'Hotels/Restaurants/Cafes' by this distribution? Would you consider these classifications as consistent with your previous definition of the customer segments?

Answer:
The distribution seems to strongly indicate that there are two clusters of customer types in this data and that they correspond to the clusters that were determined through K-Means. There are some customers who are classified in the wrong cluster, but there do seem to be two clearly defined categories in this distribution. The third sample data point that I chose originally appears to be an outlier from both clusters, but it was categorized as a “Hotel/Restaurant/Cafe” even though it is far from the centers of both clusters and was originally predicted to be in Cluster 1.

I would expect the concentrated mass of points at the center of the “Retailer” cluster to correspond to customers that would be considered purely belonging to that category. Similarly, if a point is closer to the center of the “Hotel/Restaurant/Cafe” cluster, I would expect it to be a pure representation of that category.

These categories are largely what I had in mind when I referred to the two groups as “grocery stores” vs. “restaurants,” but I had not considered other businesses such as hotels, and the “Retailer” category could include other types of customers besides grocery stores.